Week-6

APRIL 2022

M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Multiple Linear Regression

MONDAY
MARCH 2022
21

WEEK 13
080-285

The Multiple Regression Model (MLR/MRM)

- Multiple Regression Model (MRM): model for the association in the population b/w multiple explanatory variables and a response.
- K: the number of explanatory variables in the multiple regression (k=1 in simple regression)

~~The Multi~~

The response Y is linearly related of k explanatory variables X_1, X_2 and X_k by the equation.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \rightarrow \text{error}$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$
$$E[\varepsilon] = 0$$

- The unobserved errors in the model
1. are independent of one another
2. have equal variance, and
3. are normally distributed around the regression ~~expression~~, equation.

- While the Simple Linear Regression Model (~~SRM~~ SLR) bundles all but on explanatory variable into the error term, multiple ~~regression~~ regression allows for the ~~regression~~ inclusion of several variables in the model.

  e.g. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ → Marginal slope

- In the ~~SRM~~ MLR, residuals departing from normality may suggest that an important explanatory variable has been committed.

  e.g. $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k X_k + \varepsilon$ → Partial slope

Interpreting Multiple Regression

- R-squared and S_e → Std. deviation of error term (σ^2)

- $\bar{R}^2$ is known as the adjusted R-squared. It adjusts for both sample size n, and model size k. It always smaller than R^2.

- The residual degrees of freedom $(n-k-1)$ is the divisor of Se. $\bar{R}^2$ and Se move in opposite direction when an explanatory variable is added to the model ($\bar{R}^2$ goes up while Se goes down)

• Calibration Plot

• Calibration plot: scatterplot of the ~~ro~~ response Y on the fitted values $\hat{Y}$.

• R is the correlation b/w $\hat{Y}$ and Y; the tighter data cluster along the diagonal line in the calibration plot, the larger the R value.

• Marginal and Partial Slopes

Marginal Slope:- The change in Y variable with 1 unit change in X variable.

• Partial Slope: Slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.

• Marginal Slope: slope of an explanatory variable in a simple regression.

• Path Diagram

• Schematic drawing of the relationships among the explanatory variables and the response.

• Collinearity: very high correlation among the explanatory variables that make the estimates in multiple regression

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

uninterpretable.

8

Examples

9

10 Significance F = p-value
Coefficients = β_1 → Marginal Slope
11 Residual = Total − Regression (Degree of Freedom)

12 H_0 if $\beta_1 = 0$
H_a if $\beta_1 \neq 0$

1

2

3

4

5

6

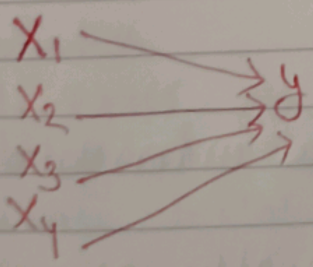
Collinearity

• Variance Inflation Factor (VIF)

• Variance inflation factor: quantifies the amount of unique variation in each explanatory variable and measures the effect of collinearity.

• The VIF for X_j is $VIF(X_j) = \frac{1}{1-R_j^2}$

where R_j^2 is the Coefficient of Determination in the regression of X_j on ALL of the other explanatory variables.



$$VIF(X_1) = \frac{1}{1-R_1^2}$$

$(R_1)^2$: Regression where, X_1 = Response Variable

X_2, X_3, X_4 = Explanatory variable

$$VIF(X_2) = \frac{1}{1-R_2^2}$$

R_2^2: Regression where, X_2 = Response variable
X_1, X_3, X_4 = Explanatory variable

Why does VIF matter?

- The standard error in estimation of the partial slope gets inflated to VIF.
- Typically,

$$se(b_1) = \frac{S_e}{\sqrt{n}} \times \frac{1}{S_x}$$

Standard error

Std. error in error term

Std. dev. of particular explanatory variable

- With VIF

$$se(b_1) = \frac{S_e}{\sqrt{n}} \times \frac{1}{S_x} \times \sqrt{VIF(X_1)}$$

VIF

- If the explanatory variables are uncorrelated, then $R_j^2 = 0$, and VIF = 1
- However, if the explanatory variables are correlated then VIF > 1. Larger the VIF, larger is collinearity.

- Large VIF also substantially increases the standard error in predicting the partial slopes (se(b)). Thereby making those predictions unreliable

- Inference for One Coefficient

- The t-statistic is used to test each slope using the null hypothesis $H_0: \beta_j = 0$.
- The t-statistic is calculated as

$$t_j = \frac{b_j - 0}{Se(b_j)}$$

- Signs of Colliniarity

- R^2 increases less than we had expect.
- Slopes of correlated explanatory variables in the model change dramatically.
- The F-statistic is more impressive than individual t-statistics.
- Std. error for partial slopes are larger than those for marginal slopes.
- Variance inflation factors increase.

- Remedies for collinearity

- Remove redundant explanatory variables
- Re-express explanatory variables.
- Do nothing if the explanatory variables are significant with sensible estimates.

TA Session

Adjusted R^2 = $$1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)}\right]$$

8

9 <u>Multicollinearity</u> :- Need three correlated variables.
Three or more variables are correlated to each other.

10

F- Statistic

11

$$F = \frac{R^2}{1-R^2} \times \frac{n-k-1}{k}$$

12

1

2